The Statistician's Concern With Experiments

by

George W. Snedecor, Iowa State College*

It seems odd to me that a statistician should be concerned with experiments. I suspect that you have much the same feeling. Many of you are scientists who look on the experiment as your own essential tool. What business has a statistician to assume authority toward the characteristic feature of your profession? It is my purpose this evening to discuss the part played by statistics in experimentation.

Historically this is not difficult. When Sir John Russell added Dr. R. A. Fisher to his staff at Rothamsted, he expected him to elicit additional information from the data which had accumulated over the years. Fisher soon detected gaps between the experimental procedures then in vogue and the statistical methods which were assumed to apply. In order to bridge the gaps he elaborated de novo a mathematical theory of experimental design together with the necessary statistical methods to implement it. Under Fisher, statistics ceased to be a more or less empirical device for processing experimental data and became established as an integral part of the scientific method.

Statistics enters scientific method at the point where observations are compared with the hypothesis they were designed to test. Now there is a profound gulf between hypothesis and experimental observations. The latter are based on only minute portions of the material with which the hypothesis deals. This would raise no problem were it not for ever-present variation. No portion of the material is exactly like any other portion and no two measurements are made under precisely the same circumstances. The environment varies in both space and time. The experimenter himself changes from moment to moment as do the instruments with which he makes his measurements. In what manner and in what degree may the facts observed in a fragment of experimental material be imputed to the aggregate comprehended by the hypothesis? The bridging of the gulf between experiment and hypothesis is precisely the problem of theoretical statistics.

---

*Presented on November 24, 1953, as a Goldwin Smith lecture. Mimeographed by the Biometrics Unit, Warren Hall, Cornell University, as BU-47-M, November 1953.

In statistical terms, an experiment is a device for drawing a sample from a population. Based on the facts observed in the sample, inferences are made about the population. This inductive process, leading to new knowledge, is logically hazardous so that conclusions must always be considered uncertain. It is the function of statistics to make a numerical evaluation of this uncertainty. Statistics, then, is that part of scientific method concerned with the drawing of conclusions from experiments.

As I have indicated, it is variation which causes the uncertainty of inference. Variation is inherent in both the experimental material and the observing mechanism. The relative amounts of variation in these two systems are worthy of comment. In some physical and chemical experiments, the materials are considered practically invariable as compared to the measuring devices, including the human link. The determination of so fundamental a constant as the speed of light is an example. This situation, leading to preoccupation with variation in the measuring system, gave rise to the theory of error or theory of observations. The opposite extreme is characteristic of the biological sciences. Consider the problem of determining the life span of the human male in birth-registration areas of the United States. Inaccuracies in measuring age at death are so small in comparison with the variation among individuals that age is often taken as known exactly. This point is emphasized in a most useful method, regression, where one of the assumptions usually made is that the independent variable is measured without error. I am not aware of any differences in the theory applicable to these two extremes or to the intermediate ranges, but many of the appropriate statistical methods are distinctive. Perhaps it should be made clear at this point that the speaker's interest and experience have been in the second environment where the taking of measurements is not ordinarily a major problem.

Variation is due to a variety of causes. For convenience such causes may be assigned to two categories, the one that results in <u>inaccuracy</u> and the other that produces <u>imprecision</u>. Causes of inaccuracy are carelessness in the conduct of the experiment, mistakes made in the recording of the results, and major environmental disturbances. The probabilities associated with these causes are unknown. On the other hand, imprecision results from great numbers of minor incidents in the

environment together with unavoidable variations in the experimental material. These random fluctuations result in errors of observation which have the following properties: (i) they are equally likely in defect or in excess of the true value, and (ii) they are more often small than large, very large errors being rare. More precisely, errors of observation are often said to follow a normal distribution with zero mean. In the design and conduct of an experiment, every precaution should be taken to reduce inaccuracy and to measure imprecision. Inaccuracies are erratic events whose effects on the outcome of the experiment cannot be assessed. But imprecision follows the known laws of probability and leads to the statistical measure of fallibility in the conclusion.

Let me illustrate the role of statistics by describing a simple type of experiment. For this I have chosen a familiar one, the organoleptic test known as the triangular. The subject is presented with three portions of some food product, say orange juice. He is told that two portions have received the same treatment but that the third was treated differently. The subject is now asked to separate the two identically treated portions from the third. If he is successful, this is evidence that the difference in treatments has produced in the product a difference that can be detected by the subject.

The first feature to be emphasized is that the experimenter had a hypothesis to be tested; the hypothesis that the difference in treatment had caused no difference in the taste of his product. Perhaps this will be more obvious if one treatment is thought of as a new one, cheaper than the old; a treatment that can be advantageously substituted for the old if the acceptability of the juice is not impaired. So, the investigator sets up what the statistician calls the null hypothesis, then proceeds with the experiment for the purpose of testing it. This you will recognize as a characteristic step in scientific method. This step is dictated by the experimenter; the statistician seizes upon it as the starting point of his theory.

The next feature to be considered is the well known requirement of scientific method that the possible outcomes of the experiment must be enumerated in advance and decisions reached as to what conclusions shall follow. In the orange juice experiment there are three ways to segregate the samples. If these samples are labelled A, B, and C, the subject may say that A and B are alike, or A and C, or B and C. One of these com-

binations indicates that the subject can distinguish the products of the two treatments and leads to rejection of the null hypothesis that there is no difference. The other two combinations lead to acceptance of the hypothesis. The consequence of acceptance is that the manufacturer will substitute the new and cheaper method for the old.

It is to be observed that elaborate precautions must be taken to protect the subject from extraneous bases of judgment. His actions must result from only the particular quality which the experiment is designed to test. If this quality is flavor, then all other qualities must be removed from observation. For example, if color could give a clue to the treatments, then filtering lights must be provided or else the subject must be blindfolded. If the subject has bases for judgment other than the contemplated quality then the experiment is worthless.

It is just here that the statistician begins to be vitally concerned. His business is to evaluate the uncertainty of the conclusion reached. In order to do this he must know the probability of acceptance and rejection because his evaluation is based on known probabilities. Now the probabilities of acceptance and rejection depend upon the probabilities of the several possible outcomes, three in the orange juice experiment. These probabilities can be known if the experimenter has taken the precautions, mentioned above, to institute proper controls. The easy way to learn the probabilities of the outcomes is to make them equal. This is readily done by randomization. If the three portions of orange juice are selected at random from the two treatments, subject only to the restriction that both must be represented, and if they are presented in random order to the subject without extraneous identifications, and finally if the null hypothesis is true, then the probability of each outcome is one-third. This means that on the assumption of the truth of the null hypothesis, the probability of accepting it is two-thirds whereas the probability of rejecting it, despite the fact that it is true, is one-third. In this way, the uncertainty of the experimenter's conclusion is evaluated; he takes a 1 in 3 chance of rejecting the null hypothesis even though it is true.

"But", you say, "what chance do I take if the null hypothesis is not true?" The answer to that question is not so easy because it depends on the magnitude of the difference between the qualities being investigated. Naturally, if this were known, there would be no experiment. But

some realistic assumptions can be made. One aspect of this question will be discussed later under the heading, "size of experiments". What you are really asking about is the sensitivity of the experiment to small differences. It is clear that the triangular test, as I have so far described it, is not sufficiently sensitive to warrant critical decisions.

This brings us to a third feature of nearly all experiments; that is, replication. There are several reasons for replication but our present interest is to decrease the risk of a false conclusion. The triangular test may be repeated, using fresh preparations of juice, a new randomization, and perhaps a second taster chosen at random from the population of potential purchasers of the product. Suppose the two like preparations are again detected. The probability of two successes is one-third times one-third, or one-ninth, assuming the null hypothesis to be true. If the hypothesis is now rejected, the probability of error is one in nine. This is still not a great risk, the chance being only slightly less than that of 3 successive girls born into a family, not an unusual event. If the change to the new method of manufacture involves any considerable financial outlay, the producer is likely to demand more certainty before a decision is made.

Another replication with successful segregation of the samples brings the probability to one in 27; yet another, to one in 81, and so on. Somewhere along the line, the producer may decide that there really is a difference between the products. When he does, he will have the advantage of knowing exactly the hazard involved in his decision.

I think you now see the role played by statistics in experimentation-the evaluation of the uncertainty involved in the inductive process of generalizing the facts observed. Every experimenter encounters this hazard and more or less consciously estimates the uncertainty. Statistics furnishes authentic methods of estimation, leading to exact statements of the probability of false conclusions.

You will see, also, that concern with the evaluation of uncertainty leads the statistician to examine every step in the process. There must be a clear statement of the objective before the null hypothesis can be formulated. The proposed measured variable must be examined to learn if it will lead to an exact test of the hypothesis. The scheme of randomization must be checked. Then this question must be answered: is the sensitivity sufficient to detect the effects being investigated? As for

the design, the statistician will consider whether the experimental questions will be answered unambiguously. Even with the conduct of the experiment the statistician is concerned because inaccuracies may ruin his measure of imprecision.

To this point I have in the main confined my discussion to that unique feature of theoretical statistics which makes it an indispensable part of scientific method, the evaluation of the fallibility of conclusions. But, as I have just said, the statistician was led to scrutinize the whole experimental procedure including the design. One consequence was R. A. Fisher's classical book, "The Design of Experiments". It turns out that some designs are more effective than others in producing information; that is, some will yield more information than others though the cost may be the same. For a simple illustration, let us revert to the tasting experiment and consider whether a change in the design might result in more information per dollar spent.

Suppose that there are two replications of the triangular test. These require the assessment of six portions of the juices, let us say by the same subject. If correct separations have been made in both tests the probability of falsely rejecting the null hypothesis if true is one in nine. But change the design, presenting six randomized portions to the subject at one time, stating that two portions have been treated in one way and the other four in another. Now the probability of a correct separation, if the subject can really detect no difference, is only 1/15, so that there is only one chance in 15 of rejecting the null hypothesis if it is true. Accordingly, by a slight change in design and with no extra cost, the probability of a false conclusion from six portions has been decreased from 1/9 to 1/15. This means more information for the same money. I am assuming, of course, that the subject has no more difficulty in judging the six portions at a single presentation than he had when there were two batches of three.

Such cost accounting is a prominent feature of survey design. It has been listed as one of the obligations that the statistician owes to his colleagues. Let us look at an experiment where it could have been used to advantage.

This is a study of sampling technique carried out at the Georgia Agricultural Experiment Station.*

---

* Southern Cooperative Series Bulletin No. 10, March 1951.

The matter being investigated was the concentration of riboflavin in turnip leaves of different sizes and at different dates of harvest. Three sizes of leaves were used, designated as small, medium, and large. One leaf of each size was picked from each harvested plant. The process was repeated on each of 12 days of harvesting.

On each day and on each leaf-size, three riboflavin determinations were made. These determinations were rather expensive procedures. The question arises as to whether such extensive chemical analyses were justified. It seems possible that some estimates might have been made in advance of the experiment, but we can do no more than look at the results. The analysis of variance is as follows:

Source of Variation	Degrees of freedom	Mean Square	Estimates of Components of Variance
Replications or Days	11	1133.30	σ_R^2 : 119.33
Sizes of Leaves	2	383.06	
Experimental Error	22	59.30	σ_E^2 : 17.20
Determinations	72	7.69	σ^2 : 7.69

It is clear that the big source of variation in riboflavin concentration is from day to day. To evaluate the mean precisely, many daily harvests are necessary, a fact that was anticipated by the investigators. On the other hand, the variation among determinations is relatively small. This suggests that the cost of this experiment might have been decreased by making fewer determinations, and that little or no information would have been lost. Let us examine the evidence.

The variance of a treatment (or leaf size) mean is

$$s_{\bar{x}}^2 = \frac{s^2 + Ds_E^2}{RD}$$

where D is the number of determinations and R is the number of replications or days of harvest. The information in the treatment mean is taken to be inversely proportional to the variance; that is,

$$I = \frac{RD}{s^2 + Ds_E^2}$$

Now the cost of the experiment may be said to be made up of the cost of running the required number of riboflavin determinations plus the cost

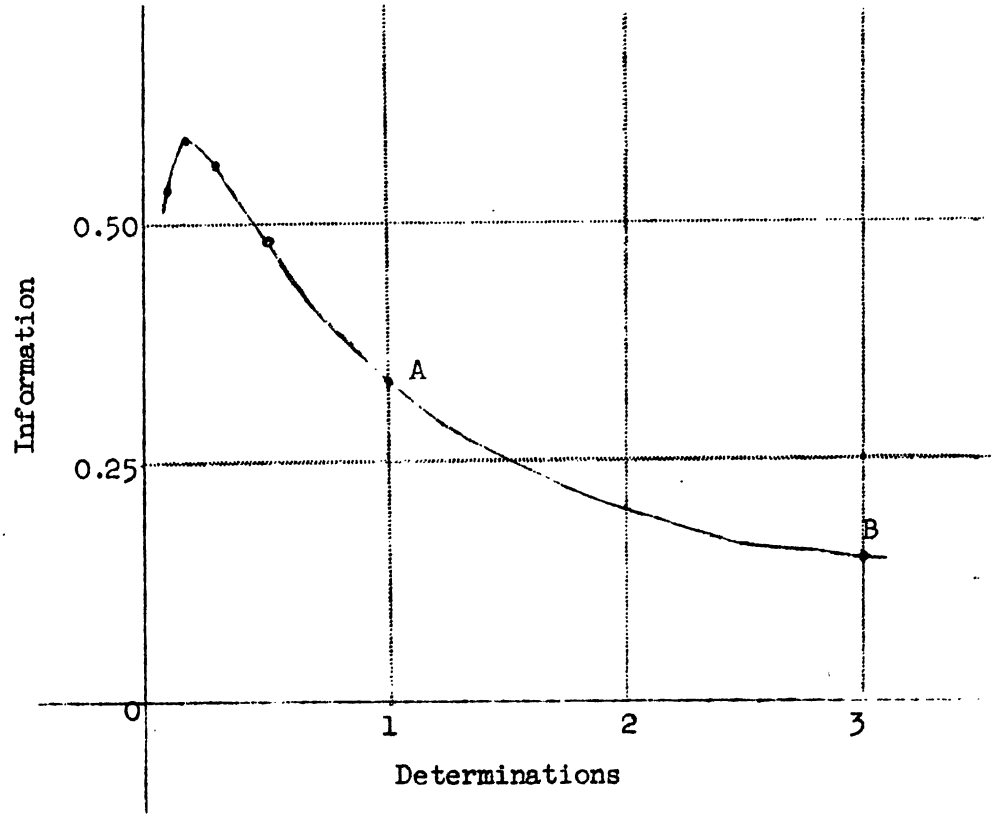
of raising and harvesting the turnip greens. If the latter is one unit per replication or day of harvest, and the former is c units per determination, then the total cost is

$$C = R + RDc = R(1 + Dc)$$

Now I cannot calculate costs exactly but I know that in Georgia turnip greens are not expensive. Allowing something for the extra care and necessary supervision in an experimental setup, perhaps 25¢ is the upper limit on the cost of the greens for a day's run. As for the cost of a riboflavin determination, I learned that our chemical service would provide this for a minimum of $3.00. Putting all these data together, the formula for information becomes

$$I = \frac{111\ D}{(7.69 + 17.10D)(1 + 12D)}$$

The number 111 is the dollars paid for the experiment as it was designed; 12 is the ratio of the cost of a determination to that of a day's supply of turnip greens. The graph of this equation is given in the figure.



Information available at original cost with varying numbers of determinations of riboflavin.

The point B shows that the information supplied by the experiment was 0.152 units. The point A indicates that, for the same cost, 0.342 units of information might have been had by doing only one determination instead of three on each batch of greens, an increase of 125%. Putting it another way, the information got from the original experiment could have been bought for $48.75, only 44% of the actual cost.

As I suggested earlier, hindsight is easy whereas it is foresight that makes for efficiency. In order to improve the design I had to draw on the original experiment for necessary information. To what extent this might have been anticipated I do not know. But of one thing I am certain: after the experiment had been under way for a few days, a preliminary statistical analysis would have revealed the facts clearly. At that time a few cents spent on calculation would have saved dollars for the experiment station.

Other features of cost accounting are two of the tests which the statistician applies to new experimental designs: (i) is it easy to repair damages that may occur during the progress of the experiment; and (ii) is it cheap to summarize the data? Most designs currently in use require only modest amounts of calculation for working up the results. From among possible designs furnishing similar amounts of information, that one is chosen which involves the least computation. But experimenters are ingenious in thinking up new designs, not in the book. They are sometimes not so ingenious in writing down the computational procedure before the experiment is initiated. Any novel design, for which the appropriate statistical methods are not available, should be checked with a statistician in advance of the experiment to make sure that the information wrought into the data can be fully extracted at reasonable cost. That this is not often done is a commonplace in the statistician's experience. The experimenter usually comes to him after the deed is done, then seems surprised that the statistician cannot offhand provide a method of analysis.

I have in mind a complex experiment designed for three replications. The experimenter had some extra material and, having the usual enthusiasm of the researcher, decided to put in four replications. To his astonishment, he learned - after the experiment was done - that there was no known method of extracting the information from the extra data. In order to utilize the fourth replication he would have to hire a mathematical statistician for a thousand or two dollars, and this sum was not in his budget. Fortunately, the blunder came to the attention of a professor on the lookout for a thesis

topic for one of his graduate students. After a couple of years of research supplemented by consultations with various experts, a suitable method of analysis was found with the result that both the experimenter and the statistician were pleased. I am sorry to say that tampering with a complex design does not always turn out so happily.

In addition to cheapness of computation, I mentioned ease of repair as one of the desirable features of experimental designs. Despite the best efforts of the investigator, experiments sometimes go wrong. Data turn up missing for one reason or another. This may not be disastrous because there are available methods of treating the remaining data. The agronomist is familiar with the missing plot technique whereby all of the remaining information can be had at little cost of calculation.

Occasionally the repairs are found to be more costly than the worth of the desired information. Only last spring I had an inquiry from a plant breeder in Hawaii asking about the analysis of a rectangular lattice experiment in which some of the plants had died. Originally there were four papaya plants per plot, but in some of them one or two had failed to survive. In this case, the effects of decreased competition might have ruined the experiment, but the investigator had provided against this by replacing the missing plants with new plantings. Still he did not consider it valid to use the measurements on these imported plants. The result was unequal weightings in the plot data. The mathematicians assured me that an exact analysis, were it possible, would cost many hundreds of dollars. I recommended, therefore, that the investigator use the unweighted plot means in the standard form of calculation. I believe that he will lose little information by doing so, far less expensive than the cost of the exact solution.

Not infrequently one encounters experiments whose results are almost devoid of information. Some people seem to think that a statistician is a kind of magician who can conjure information from data by manipulation of a calculating machine. The statistician is not a party to this thinking. He knows that his best is to extract all the information that has been built into the data by the skill and care of the experimenter. If the design of the experiment is faulty, or if it has been sloppily conducted, the statistician can be of ltttle or no help.

I was once confronted with an extensive series of experiments at one time in vogue in Iowa. I was told that the design had been brought from the Rothamsted Experimental Station in about 1910. It was a factorial

design with two sets of fertilizers, organic and chemical. The field plan was like this:

Manure	Manure Lime	Manure Lime Rock Phosphate	Manure Lime Acid Phosphate	Manure Lime Commercial Fertilizer
Crop Residue	Crop Residue Lime	Crop Residue Lime Rock Phosphate	Crop Residue Lime Acid Phosphate	Crop Residue Lime Commercial Fertilizer

There was no replication in any one field.

In a single experiment of this kind there is no exact information. The effects of treatments and of soil variation are confounded. There is no provision for estimating treatment differences and none for evaluating uncertainty. No conclusions can be drawn from the results in any one field.

Now, these experiments were repeated in several fields having the same soil type so that there was essential replication. But there was no randomization. This means that a separate estimate of error had to be calculated for each kind of comparison, making statistical analysis expensive. Such an experiment is not altogether worthless, but it is inefficient; the cost per unit of information is excessive.

In addition to cost accounting, one may list another obligation of statisticians, the estimation of sample size. When I first became interested in statistics some 40 years ago, the question most often asked was, "How many replications must I provide?" Apparently, there has since been a good deal of stabilization of experimental customs because I seldom hear this question now. It has been transferred to the survey designers. But in new fields of investigation and, I think, in many of the older ones, consideraiion of size is a vital part of planning. During the last few years ample facilities have been provided for giving definite answers to this problem of experiment size.*

This doesn't mean that the investigator can merely look in a book and read from a table the number of replications to be used. He will find

---

* John W. Tukey: "The Problem of Multiple Comparisons", American Statistical Association Annual Meeting, Chicago, 1952

* Henry Scheffé: "A Method for Judging all Contrasts in the Analysis of Variance", Biometrika 40:87, 1953

it necessary to supply a number of facts and estimates about the prospective design and experimental material. These are as follows:

1. The type of treatment contemplated; for example, randomized blocks.
2. The number of treatments to be tried.
3. The magnitude of the difference to be detected.
4. The level of significance desired.
5. The assurance demanded that the experiment provide the answers specified.
6. An estimate of the experimental error.

The last is the most elusive piece of information. Most often it is furnished by a previous experiment or a series of them. Harris, et al.,* have suggested a way of estimating it by using a general knowledge of range in the experimental material. The estimate may be in specific units of measurement or may be expressed as a percentage of the mean. It must be accompanied by the degrees of freedom on which the estimate is based.

For an example I have taken data from an investigation of oxygen consumed in household tasks, a report from your College of Home Economics.** I have used only three of the moderately heavy activities, chiefly because of their uniformity in variance. The activities with the corresponding average oxygen consumptions (increases over standing) are these:

1. Reaching to a height of 72 inches, 109 cc./min.
2. Reaching to a height of 22 inches, 128 cc./min.
3. Reaching to a height of 36 inches, then pivoting 90°, 97 cc./min.

During each of four periods these motions were repeated 22 times per minute for four minutes. Nine subjects participated in the trials.

From the data reported I computed this analysis of variance:

Source of Variation	Degrees of Freedom	Mean Square
Subjects	8	1744
Activities	2	2195
Experimental Error	16	813
F = 2195/813 = 2.70	P = 0.10	C = 28.52/111.2 = 25.6%

*Marilyn Harris, D. G. Horvitz and A. M. Mood: "On the Determination of Sample Sizes in Designing Experiments", Journal of the American Statistical Association, 43:391, 1948

**Esther Crew Bratton: "Oxygen Consumed in Household Tasks", Cornell University Agricultural Experiment Station, Bulletin No. 873, August 1951.

As the investigator indicated, the variation was large, the coefficient of variation for these three activities being 25.6%. If there is really no difference at all among the consumptions of oxygen for these three activities, there is yet one chance in ten of having an F of 2.70 or more turned up by the experiment. In this situation a mean difference would have to be at least 34 cc./min. to be accounted significant at the 5% level; even the largest of those reported is not so great. Notice too that 34 is more than 30% of the mean for these treatments so the experiment cannot be accounted sensitive.

Now in reviewing these results, let us assume that this fact emerges: If any difference between oxygen consumption for these activities is really as large as 20 cc. per minute, it is essential to know it. How many subjects (that is, replications) are necessary to detect so small a difference at the 5% level? It is specified also that the investigator is willing to run no larger chance than one in five of being unsuccessful in the new endeavor.

We now have all the data necessary for estimating sample size. A randomized blocks experiment with three treatments is planned to detect any difference as large as 20 cc. per minute, the 5% level of significance to be used. An estimate of error, $s^2 = 813$ with 16 degrees of freedom, is available. The experiment must be large enough to allow only a moderate chance (one-in-five) of failure. Using the method devised by Dr. John W. Tukey, I find that 18 subjects will be required, twice the original number. If the larger experiment is tried and differences of 20 or more do not turn up, the investigator may be reasonably sure that they do not exist in the population. Putting it another way, if differences as large as 20 cc./min do exist in the population, there is reasonable assurance that they will be detected by the new experiment.

Tukey's method requires only a few minutes of calculation. There seems to be no reason why any researcher should not apply it before deciding on the size of his experiment. The computations will show him either the size required to detect a specified difference or the difference he can expect to detect with a given number of replications. If necessary data are not available in advance, the technique may be applied even after the experiment is under way as soon as preliminary estimates of error are accumulated.

When sample size is routinely estimated as part of experimental projects, inconclusive results will be far less common than they are now. Moreover, many contemplated experiments will be abandoned in advance with consequent saving of time and money because it will be obvious that no worthwhile decisions can be reached with the available resources. More crucial experiments will be planned with fore-knowledge that many replications will be required, extending perhaps over a long time or to various locations.

In summary: The statistician is primarily concerned with the inductive conclusion from an experiment. This involves uncertainty, an uncertainty which under ideal conditions can be exactly evaluated. Such evaluation is the fundamental problem of scientific statistics. But in seeking to insure a correct evaluation of uncertainty, the statistician becomes interested in all the accessories of the process: the design, the size, and even the conduct of the experiment. Moreover, after the experimental work is done, the concern of the statistician continues. There are sometimes repairs to be made and there are questions about sample size and about economy in the design of prospective experiments. Altogether, I think you will agree that the statistician's concern is not only with the evaluation of uncertainty but includes every step in the experimental procedure, from planning to post-mortem.